63-2-4

296 598

MEMORANDUM
RM-3271-PR
FEBRUARY 1963

REMOVING THE NOISE FROM THE QUANTIZATION PROCESS BY DITHERING: LINEARIZATION

G. G. Furman

PREPARED FOR:
UNITED STATES AIR FORCE PROJECT RAND

The RAND Corporation
SANTA MONICA • CALIFORNIA

MEMORANDUM
RM-3271-PR
FEBRUARY 1963

REMOVING THE NOISE FROM THE QUANTIZATION PROCESS BY DITHERING: LINEARIZATION

G. G. Furman

This research is sponsored by the United States Air Force under Project RAND — Contract No. AF 49(638)-700 — monitored by the Directorate of Development Planning, Deputy Chief of Staff, Research and Technology, Hq USAF. Views or conclusions contained in this Memorandum should not be interpreted as representing the official opinion or policy of the United States Air Force. Permission to quote from or reproduce portions of this Memorandum must be obtained from The RAND Corporation.

The RAND Corporation
1700 MAIN ST. • SANTA MONICA • CALIFORNIA

PREFACE

In processes involving analog machines, digital machines, or a combination of these, the accurate transmission of information is of great importance. This is sometimes aided by the application of independent quantizer activators called dithers. The present Memorandum, written by a RAND Corporation consultant, analyzes questions concerning dithers and gives some quantitative evaluations regarding them.

The author expresses thanks to Professor Charles Susskind, who read the manuscript and made editorial suggestions.

SUMMARY

This Memorandum treats the linearization of the highly important multistep quantizer nonlinearity by the application of independent quantizer activators called dithers. For the dithered quantizer acting (as it often does) in conjunction with a low-pass or band-pass filter, numerical answers are given for the first time to the questions: (a) What is the equivalent quantizer gain? (b) What upper bounds does the dither place upon the maximum deviation from linearity? (c) How does one determine, for given specifications, a dither amplitude so that the system is optimally dithered? Such information, with regard to two time-periodic dithers (the sinusoid and sawtooth) makes it possible to effect analog-to-digital-to-analog conversion (to name one application) with no apparent loss of information even when the quantization is rough. The properties of the sinusoidal and sawtooth waves as quantizer linearizers are developed in detail; it is shown that the sawtooth is superior to the more popular sinusoid in most important respects.

CONTENTS

REMOVING THE NOISE FROM THE QUANTIZATION PROCESS BY DITHERING: LINEARIZATION

1. INTRODUCTION

The operation of amplitude quantization, which transforms continuous-amplitude variables into variables whose amplitudes can have only discrete levels, is being incorporated in a growing number of control and computer processes. In general any periodic quantization type of nonlinearity can be represented by the staircase characteristic of Fig. 1; but the ordinary rounding-off process, which this paper treats, requires that the quantizer characteristic be made square and symmetrical, i.e., with quanta size $q = \ell_h = \ell_b$ and bias $h = v = 0$. In the present paper, such rounding-off transducers (for example, those that round off to the nearest integer, with $q = 1$) will simply be called "quantizers," and their presence will be indicated by the symbol Q. Note that it is possible to transform the quantizer nonlinearity into the more general nonlinearity of Fig. 1 if the linear operations of summation and pure amplification are allowed both to precede and to follow the quantization operation itself [1, 2]. Therefore the results obtained here can be readily extended to other staircase nonlinearities.

Although it is evident that digital computers must entail amplitude quantization, many other nonlinear processes that are apparently unrelated to quantization are in fact

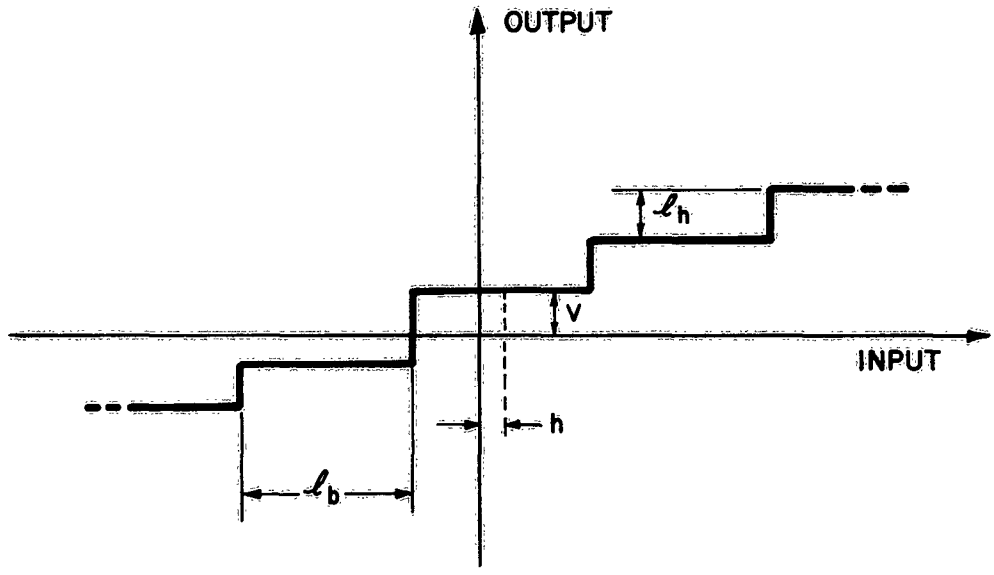

Fig. 1

reducible to processes that are linear except for a quantizer. For example, a pulse-frequency modulator, which emits identical narrow pulses at a rate proportional to its input, is equivalent to a transducer that subjects its input to integration, quantization, and differentiation, in that order.

Amplitude quantization is now a popular means of preserving signal accuracy in the face of transmission noise. Moreover, the use of digital computers in otherwise nonquantizing systems is growing with the versatility and sophistication of the computer sciences [3].

The stimulation of such quantizing systems by externally generated signals (dithers) holds great promise [1] as a means of doing away with the two main disadvantages of quantization: (a) the information loss inherent in the rounding-off process, and (b) the possibility of limit-cycle oscillation. At the same time the advantages of digital computation and digital data transmission can be fully exploited.

The idea of linearizing highly nonlinear transducers by means of an externally generated signal is not a new one. McColl [4] and Lozier [5] both recommend that a sinusoidal dither be applied to a motor drive system incorporating a relay, and Loeb [6] has suggested that <u>any</u> nonlinear system can be linearized in this manner. Other workers have been concerned with the problems of

stability [7] and noise in amplitude-quantizing sampled-data systems [8, 9]. Although the effects of dither on system linearization have been described quantitatively for a number of nonlinearities, for the quantizer they are here described for the first time. Some of the effects of dither on the quantizer as an operator on the statistical properties of its inputs (such as their mean square) have been reported [1, 2, and 10]. Such information is important for other reasons, but it leaves the question of linearity open.

2. PROBLEM FORMULATION

We now formulate the present problem, as schematized in Fig. 2.

A. The following information is given:

a. Transfer functions A and C may be linear or nonlinear, and sampled data or time continuous, but they are otherwise unspecified. Hence we may have $A = 1$, $C = 0$.

b. Transfer function B is linear and has either a low-pass or band-pass frequency characteristic. Also, if B is sampled data, the sampling rate is high relative to the frequencies of $o_q(t)$, so that for purposes of analysis B may be approximated by a continuous transfer function.

c. The dither may be injected at the quantizer input (as in Fig. 2) or, if that is not possible and A is linear, at the system input.

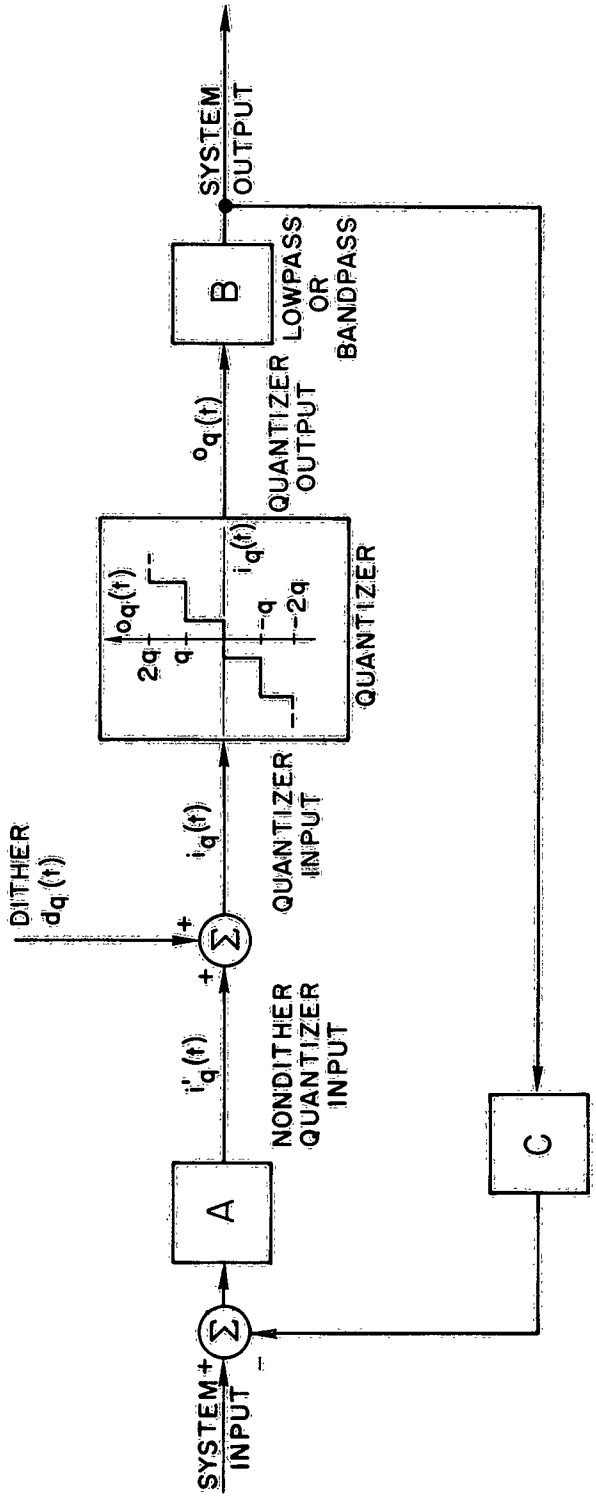
SYSTEM INPUT
Σ
+
-
A
$i'_q(t)$
NONDITHER QUANTIZER INPUT
DITHER $d_q(t)$
Σ
+
+
$i_q(t)$
QUANTIZER INPUT
$o_q(t)$
2q
q
-q
-2q
$i_q(t)$
QUANTIZER
$o_q(t)$
QUANTIZER OUTPUT
B
LOWPASS OR BANDPASS
SYSTEM OUTPUT
C

Fig. 2

d. The notation is:

$i_q'(t)$ = nondither quantizer input,

$d_q(t)$ = injected dither,

$i_q(t)$ = total quantizer input = $i_q'(t) + d_q(t)$,

$o_q(t)$ = quantizer output = $i_q(t)$ rounded off to the nearest quantum.

B. The designer is to select a dither type so that with regard to its input-output relationship the dithered system behaves as if the dither were not present (i.e., as if $d_q(t) = 0$) and the quantizer had a pure gain of unity. The quantizer is then said to be linearized.

C. The designer works under the following restrictions:

a. The instrumentation for a dither source should not appreciably increase the complexity of the system.

b. Even though it is sometimes advantageous from a theoretical viewpoint to employ dither amplitudes that range over many quanta, for reasons of economy the designer must not allow the dither to increase the dynamic range of the quantizer appreciably.

Expressed in other terms, any difference between the quantizer input and output is undesirable in the sense that it can represent an information loss. For this reason it is called a "rounding-off error" or "quantization noise" $n_q(t)$ [11], with $n_q(t) = o_q(t) - i_q(t)$. The maximum magnitudes of n_q occur for values of i_q which are halfway between adjacent quanta levels. Hence we have $-q/2 \leq n_q \leq q/2$.

Dither, when used effectively, does not cause n_q to vanish. Rather, high-frequency dither is employed to drive the noise-frequency spectrum to regions where the noise is absorbed by B together with the dither. That is, of the total quantizer output $o_q = i_q' + d_q + n_q$, only the first term appears (as a result of its lying within the passband of B) at the system output in appreciable magnitudes.

We shall comment only in passing on the eradication of the bounded periodic oscillations or limit cycles to which quantizing systems are subject, merely to say that these oscillations are brought under control at the same time that the linearity of the quantizer is improved by dithering. The problem of controlling the amplitude and frequency of such limit cycles by means of dither (often referred to as signal stabilization) has received considerable attention since the publication of Oldenburger's exploratory paper [12].

The usefulness of a random dither having a Gaussian probability density [2, 13] as well as that of a sinusoidal dither [14, 15] in bringing limit cycles under control has been investigated. A means of controlling limit-cycle amplitudes by the manipulation of gain parameters within the loop has also been worked out [16]. Only one of these studies [2] treats the multistep quantizer, which is our present concern, but they do consider a switch or limiter nonlinearity, which is precisely what the multistep non-

linearity is reduced to when the system operates in a low-amplitude limit-cycle mode.

Before discussing the action of dither, we shall indicate why other methods of quantizer linearization may be less attractive.

Suppose, for example, that one wishes to reduce a maximum round-off error from 0.5 to 0.1 units without recourse to dither. This requires that q = 1 is replaced by q = 0.2 units. Such a change is equivalent to subjecting $i_q(t)$ to an amplification of 5 and $o_q(t)$ to an amplification of 1/5; q = 1 is retained. It is evident that the dynamic range of the quantizer has been increased five-fold, requiring a greater capacity. Should a new quantizer with q = 0.2 units be available, it will certainly be more complex than the original one. Of course such an exchange of quantizers may be physically impossible in the first place because the operation of quantization may not be carried out in a separate component; it has been isolated in the schematic only for the purposes of analysis. Moreover, regardless of how small q is made, the quantizer continues to exhibit an inactive region or dead zone for $-q/2 < i_q' = i_q < q/2$.

On the other hand, it is possible to achieve <u>perfect</u> linearization economically by employing a certain low-amplitude dither, as we shall demonstrate.

3. DITHER TYPES

In the past, the term "dither" has become almost synonomous with "sinusoidal dither", so popular was this wave type as a linearizer. More recently a periodic, zero-average sawtooth dither has been recommended [1]. Moreover, it has been suggested [2] that a random dither such as a Gaussian one may be suitable for linearizing the quantizer. Even if the Gaussian dither should be subjected to preliminary linear, high-pass filtering, however, there is at best a nonunity probability that the dither will produce acceptable linearization. Also, there is a need of overdesigning to compensate for the relatively high probability of the dither amplitude being in the neighborhood of zero. The Gaussian dither does, however, possess certain properties that greatly simplify an analysis of the statistical properties of dithered systems; for a discussion of this point, see [10].

It is interesting to note that a comparison of the sinusoidal, sawtooth, and Gaussian dithers has been carried out experimentally [17] on a simple contactor (which is an amplitude-saturating quantizer) system, with the result that perfect linearization of the system response was achieved by both the sawtooth and sinusoid, as one would expect. On the other hand, a Gaussian dither which had been high-pass filtered produced less satisfactory results; although this random dither did help to arrest large-

amplitude limit-cycle oscillations, the system output still showed a random jitter.

At the present time two periodic dithers, the sinusoid and the sawtooth, appear to be the most suitable dither types for effecting linearization: the sinusoid because it is so easy to generate and inject, and the sawtooth because it is unique as a dither in being able to bring about complete linearization even when its peak-to-peak amplitude is only one quantum.

A. Sinusoidal Dither

Suppose that a sinusoidal dither $d_{\sim}(t)$ is injected so that

$$(1) \qquad i_q(t) = i_q'(t) + mq \sin \omega_d t$$

where m is a dimensionless quantity giving the dither amplitude in quanta, and $\omega_d = 2\pi f_d = 2\pi/T_d$ is the dither radian frequency. Then the system output will approximate the time average

$$(2) \qquad o_{qa}(t) = (1/T_d) \int_0^{T_d} o_q(t)dt$$

for the situation in which the frequency spectrum of i_q' lies in the center of the passband of B relative to ω_d (for B bandpass), or is much nearer to the origin (for B

low-pass) than ω_d; i.e., i_q' is unattenuated relative to $d_\sim(t)$. One can create such a situation by employing sufficiently high dither frequencies. Strictly speaking, the system output will then be proportional (rather than equal) to o_{qa}', but this circumstance has no bearing on linearization, which is the problem at hand.

The plots of o_{qa} vs i_q' appearing in Fig. 3 are the result of digital-computer simulation for the situation depicted schematically in that figure. In Table 1, o_{qa}/q is given analytically for three ranges of m. The task of furnishing a complete characteristic of o_{qa} vs i_q' for each value of m is simplified considerably by virtue of certain properties of the characteristic.

(1) Periodicity. Because the quantizer characteristic itself is periodic with period q, we have

$$o_{qa}(i_q' + nq) = o_{qa}(i_q') + nq \tag{3}$$

for all m, where n is an integer.

(2) Symmetry. Because $d_\sim(t) = d_\sim(T_d - t)$, we have

$$o_{qa}(i_q') = (n - 1)q - o_{qa}\,[(n - 1)q - i_q'] \tag{4}$$

for $0.5q(n - 2) < i_q' < 0.5q(n - 1)$ for all m, where n is an integer.

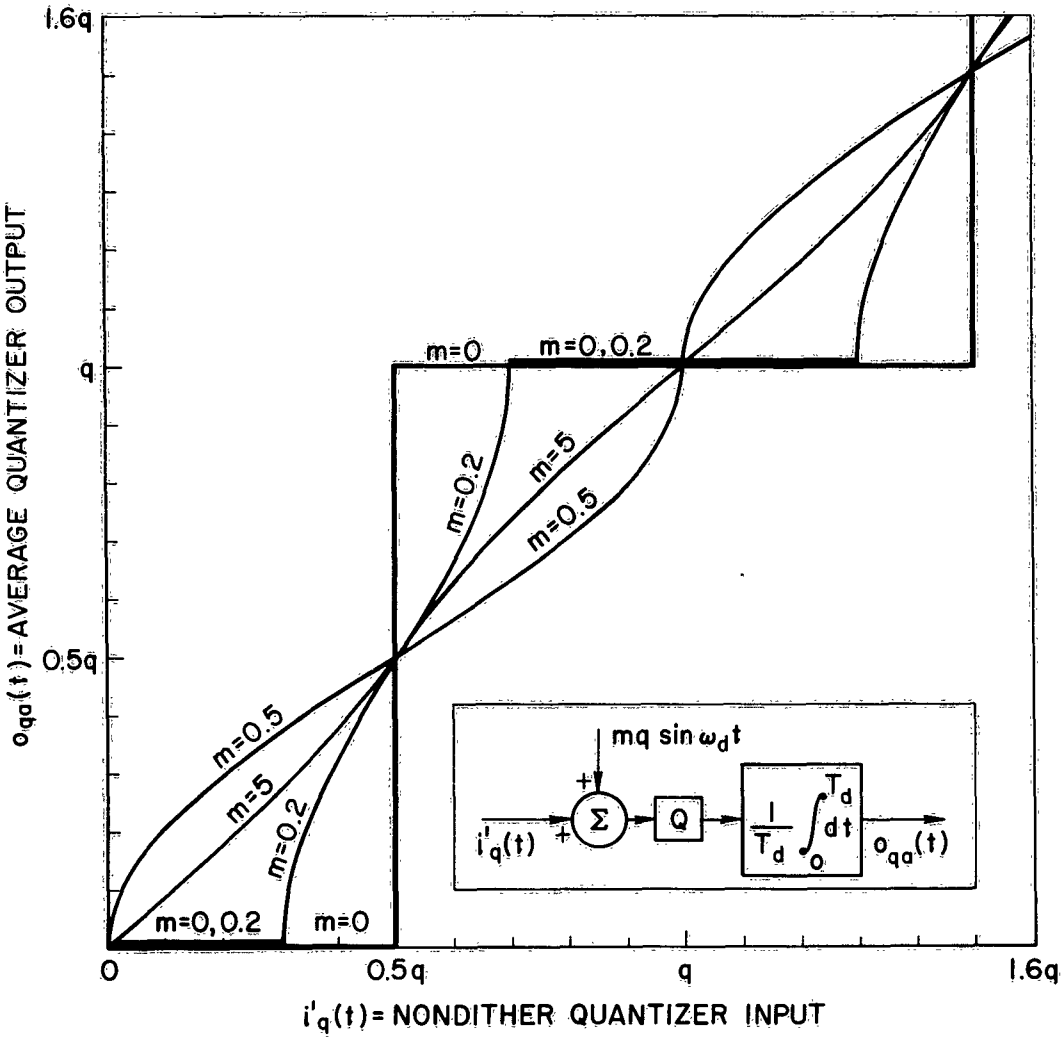
1.6q
q
0.5q
0
0.5q
q
1.6q
$o_{qa}(t)$ = AVERAGE QUANTIZER OUTPUT
$i'_q(t)$ = NONDITHER QUANTIZER INPUT
m=0
m=0,0.2
m=0.2
m=5
m=0.5
m=0.5
m=5
m=0.2
m=0,0.2
m=0
mq sin $\omega_d t$
+
+
Σ
Q
$\frac{1}{T_d}\int_0^{T_d} dt$
$i'_q(t)$
$o_{qa}(t)$

Fig. 3

Table 1

SOME EXPRESSIONS FOR o_{qa}/q IN CLOSED FORM FOR SINUSOIDAL DITHER

Definitions: $i_n' \triangleq i_q'/q$, $s \triangleq i_n' + m$,

$\theta_1 \triangleq \sin^{-1}[(0.5 - i_n')/m]$,

$\theta_2 \triangleq -\sin^{-1}[(0.5 + i_n')/m]$ and

$\theta_3 \triangleq \sin^{-1}[(1 - i_n')/m]$.

Range of m	Range of s	Condition	o_{qa}/q
$0 \leq m \leq 0.5$	$0 < s < 0.5$		0
$0 \leq m \leq 0.5$	$0.5 < s < 1.5$	$i_n' \leq 0.5$	$0.5 - \theta_1/\pi$
$0 \leq m \leq 0.5$	$0.5 < s < 1.5$	$i_n' \geq 0.5$	$0.5 - \theta_2/\pi$
$0 \leq m \leq 0.5$	$0.5 < s < 1.5$	$i_n' - m \geq 0.5$	1
$0.5 \leq m \leq 1$	$0.5 < s < 1.5$	$i_n' - m \leq -0.5$	$-(\theta_1 + \theta_2)/\pi$
$0.5 \leq m \leq 1$	$0.5 < s < 1.5$	$i_n' - m \geq -0.5$	$0.5 - \theta_1/\pi$
$1 \leq m \leq 1.5$	$1 < s < 2$	$s \leq 1.5$	$-(\theta_1 + \theta_2)/\pi$
$1 \leq m \leq 1.5$	$1 < s < 2$	$s \geq 1.5$	$0.5 - (\theta_1 + \theta_2 + \theta_3)/\pi$

(3) Oddness. By letting n take on negative values, it follows directly from property (2) that

$$o_{qa}(i_q') = - o_{qa}(-i_q') \tag{5}$$

for all m.

(4) Completeness. It follows from repeated application of properties (1) and (2) that, for all m, $o_{qa}(i_q')$ is completely specified if it is specified for the interval $0 < i_q < 0.5q$.

Properties (1) and (2) are evident in Fig. 3, where the over-all quantizer gain $OQG = o_{qa}(i_q') / i_q'$ is seen to approach unity as m increases from 0 to 5 in three increments.

Of great importance to the engineer is the maximum excursion from linearity,

$$MEL(m) = |i_q' - o_{qa}(i_q')|_{max} \tag{6}$$

which is found for any given value of m by allowing i_q' to vary freely. The smallest positive maximizing value of $i_q' = i_{qm}'$ lies in the range $0 < i_{qm}' < 0.5q$, because of the symmetry property; i.e., we have

$$|i_{qm}' - o_{qa}(i_{qm}')| = |q - i_{qm}' - o_{qa}(q - i_{qm}')| \; . \tag{7}$$

By means of computer programming, a plot of MEL vs m has been generated for $0 \leq m \leq 6$ (Fig. 4).

B. Optimal Sinusoidal Dither

Use of Fig. 4 makes it possible to solve the following practical design problems:

(a) Economical considerations dictate that some upper bound must be placed upon the quantizer's dynamic range, and consequently on m. What is the value of $m = m_{opt}$ that minimizes MEL under the constraint that $m \leq m_{max}$?

(b) Specifications call for an upper bound on MEL. What is the smallest value of $m = m_{opt}$ for which $MEL \leq MEL_{max}$?

It is apparent that, because the derivative d(MEL)/dm can be positive, in general we have $m_{opt} \neq m_{max}$. Observe that the maxima of MEL(m) (in addition to the one at $m = 0$) occur at $m \cong 0.15 + 0.5n$, where integer $n > 1$, i.e., at $m = 0.65, 1.15, 1.65, \ldots$. The minima of MEL(m) occur at $m \cong 0.45 + 0.5(n-1)$, where n is a positive integer, i.e., the minima occur at $m = 0.45, 0.95, 1.45, \ldots$.

Employing the above information, the designer proceeds to solve problem (a) in the following fashion:

(1) Determine approximately the maximum swing of $i_q'(t)$ if the quantizer were to be replaced by a pure gain of unity.

(2) Comparing this swing to the bound on the quantizer's operating range, obtain the value of m_{max}.

(3) On Fig. 4, construct the vertical line $m = m_{max}$.

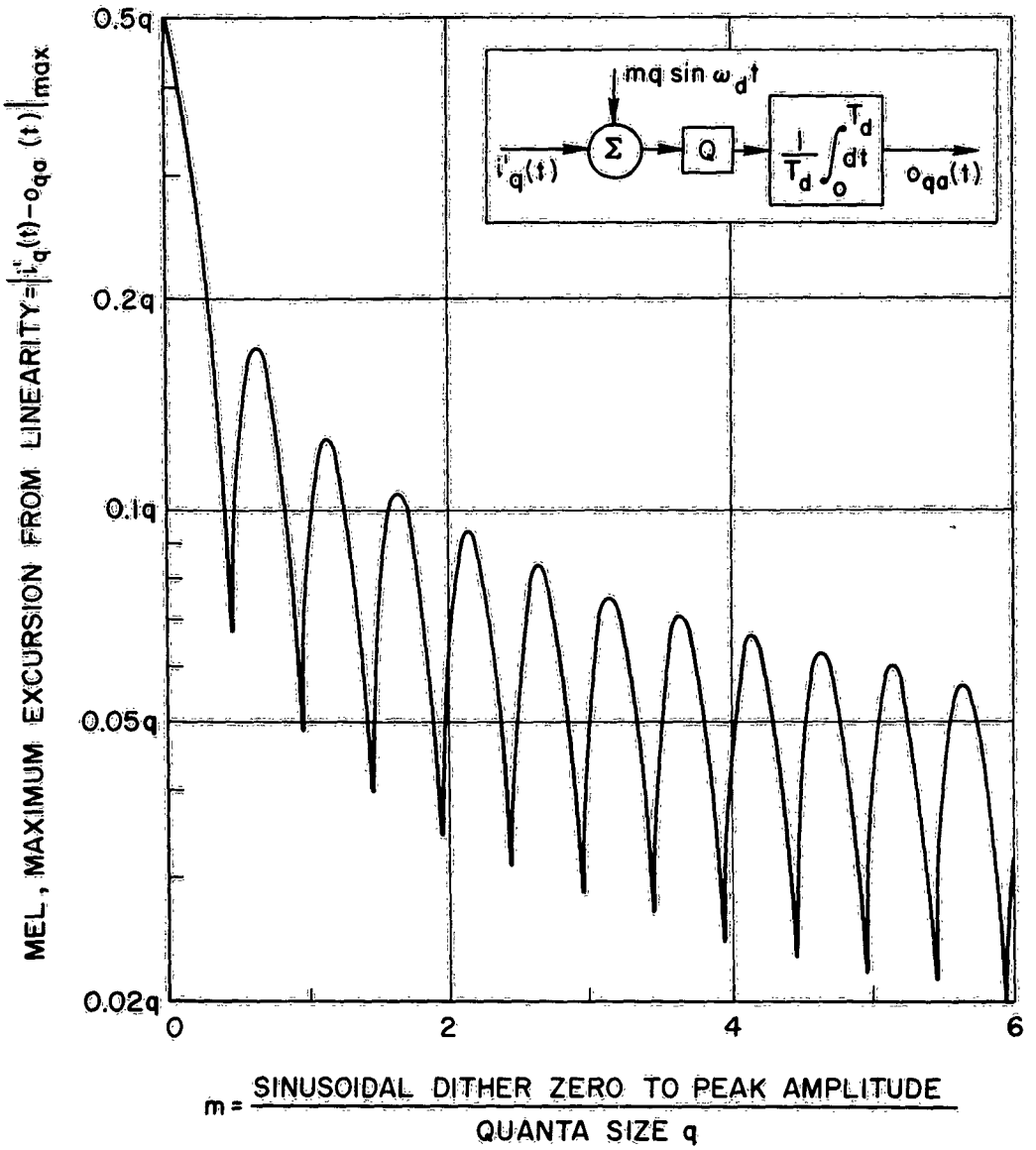
0.5q
0.2q
0.1q
0.05q
0.02q
0
2
4
6
MEL, MAXIMUM EXCURSION FROM LINEARITY = $|i'_q(t) - o_{qa}(t)|_{max}$
$m = \frac{\text{SINUSOIDAL DITHER ZERO TO PEAK AMPLITUDE}}{\text{QUANTA SIZE q}}$
$mq \sin \omega_d t$
$i'_q(t)$
Σ
Q
$\frac{1}{T_d}\int_0^{T_d} dt$
$o_{qa}(t)$

Fig. 4

(4) Take as m_{opt} that value of $m \leq m_{max}$ which minimizes MEL.

(5) Replace (theoretically) the unity gain quantizer by one having MEL = $MEL_{|m=m_{opt}}$ and repeat steps (2) through (4) until two consecutive iterations yield the same value for m_{opt}. Usually one iteration will suffice.

As an illustrative example, suppose that the i_q' swing for step (1) is $-q \leq i_q' \leq 0.9q$ and that the quantizer's operating range is $-1.5q \leq i_q \leq 1.5q$. Then $m_{max} = 0.6$, $m_{opt} = 0.45$ and $MEL_{|m=m_{opt}} = 0.068q$. If iteration yields $m_{max} \geq 0.45$, the design is complete, and the MEL has been reduced to less than 1/7 of its undithered value (0.5q) with the quantization grain unchanged.

The solution to problem (b), where MEL_{max} is specified, requires that the line MEL = MEL_{max} be constructed on Fig. 4. Proceeding along this line from left to right, choose as m_{opt} the value of m at the first intersection with the MEL-vs-m curve. For example, if $MEL_{max} = 0.05q$, then $m_{opt} = 0.95$.

C. Sawtooth Dither

Consider now the sawtooth dither $d_\Delta(t)$, periodic with period T_d, such that

$$d_\Delta(t) = mq\,[1 - (2t/T_d)] \text{ for } 0 < t < T_d \,. \tag{8}$$

The o_{qa}-vs-i_q' characteristic for the sawtooth (Fig. 5) exhibits the properties of periodicity, symmetry, oddness, and completeness ascribed above to the sinusoid. Likewise the MEL-vs-m characteristic (Fig. 6) can be obtained for $0 < i_{qm}' < 0.5q$. The sawtooth also commands additional properties:

(1) The sawtooth $OQG \triangleq OQG_\Delta$ [the subscripts $_\sim$ (or $_\Delta$) denote that the dither used is sinusoidal (or sawtooth)] is unity with $MEL_\Delta = 0$ for all $m = 0.5n$, where n is a positive integer. For a proof, see Appendix I.

(2) For $MEL \neq 0$ and $0 < i_q' < 0.5q$, the o_{qa}-vs-i_q' plot is composed of two straight-line segments of different slope passing through (0,0) and (0.5q, 0.5q), respectively. Therefore the point P at which they intersect, $(i_{qm}', i_{qm}' \pm MEL)$, completely specifies the plot. To find P for any m, note that i_{qm}' is the smallest positive value of i_q' which causes the range of $o_q(t)$ to increase or decrease (whichever calls for the smallest i_q') by one quantum from its range for $i_q' = 0$.

Consequently construction of the o_{qa}-vs-i_q' characteristic is quick and simple: For any value of $m = m_1 \leq 6$, consult Fig. 6 for the corresponding value of $MEL = MEL_1$. Strike out the integral part of m_1, retaining the part R to the right of the decimal point. If $0 < R < 0.5$, draw a straight line from the origin to point P at $[(0.5 - R)q, (0.5 - R)q - MEL_1]$. Next draw a

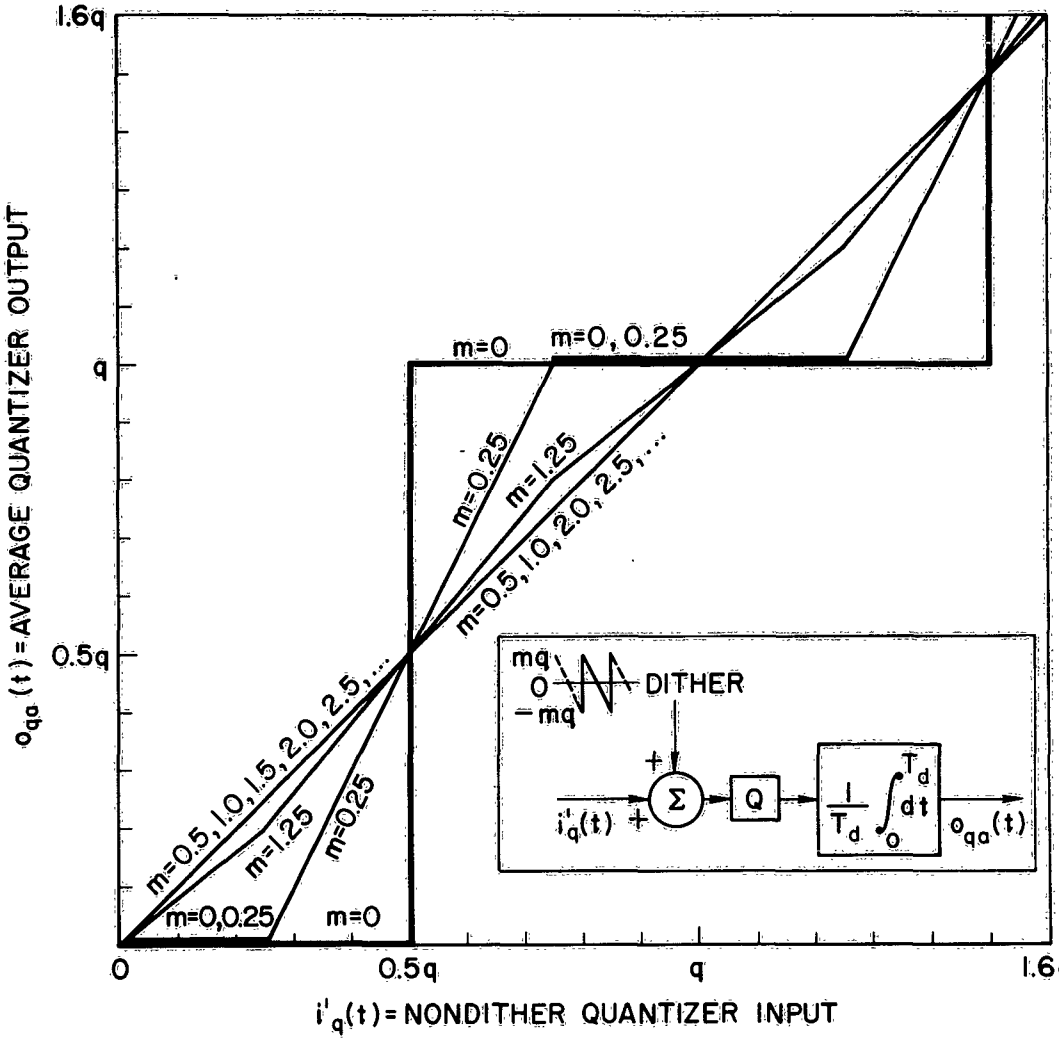
1.6q
q
0.5q
0
0.5q
q
1.6q
$o_{qa}(t)$ = AVERAGE QUANTIZER OUTPUT
$i'_q(t)$ = NONDITHER QUANTIZER INPUT
m=0
m=0, 0.25
m=0.25
m=1.25
m=0.5, 1.0, 2.0, 2.5, ...
m=0.5, 1.0, 1.5, 2.0, 2.5, ...
m=1.25
m=0.25
m=0, 0.25
m=0
mq
0
−mq
DITHER
+
$i'_q(t)$ +
Σ
Q
$\frac{1}{T_d}\int_0^{T_d} dt$
$o_{qa}(t)$

Fig. 5

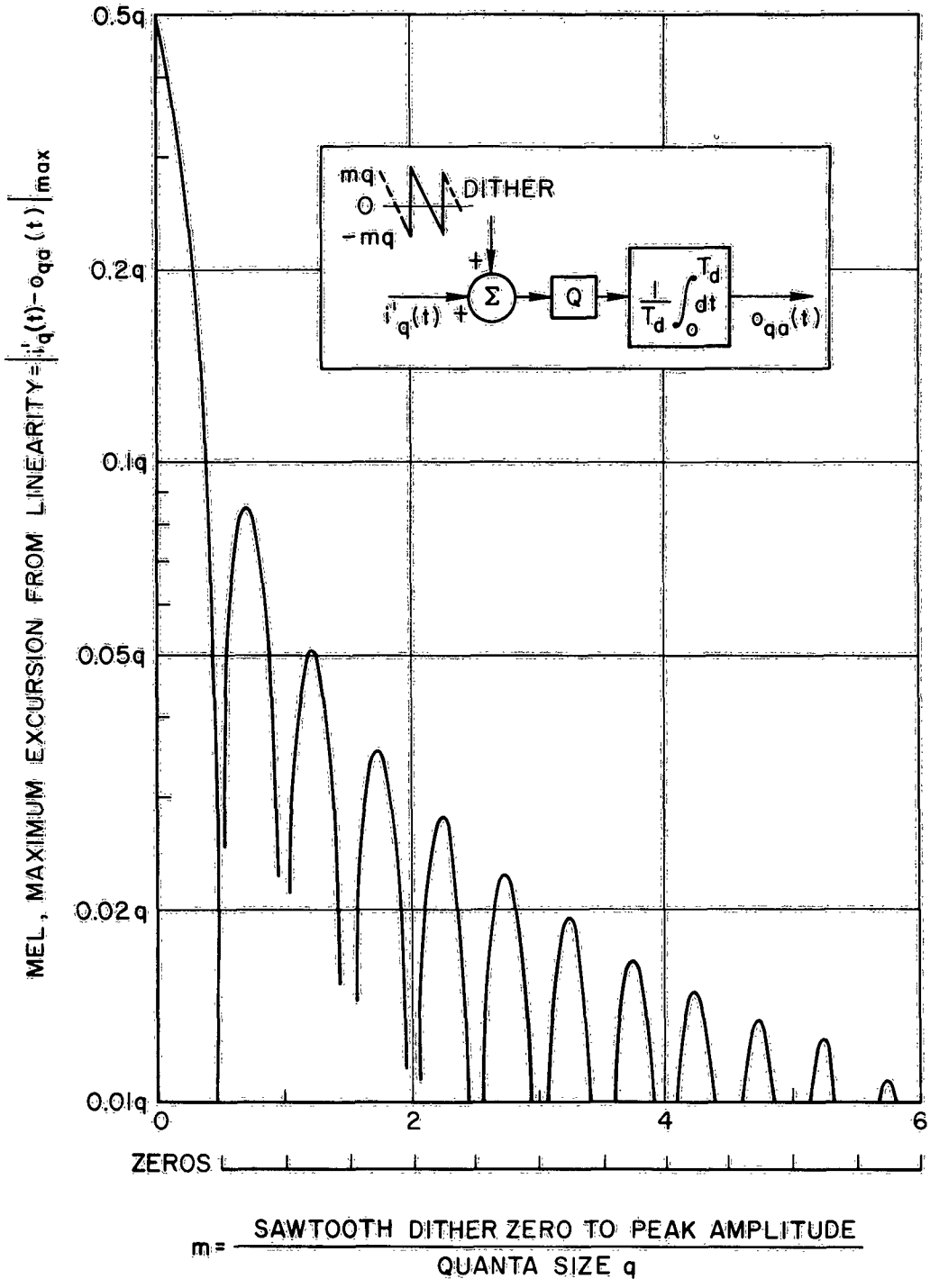
0.5q
0.2q
0.1q
0.05q
0.02q
0.01q
0
2
4
6
ZEROS
MEL, MAXIMUM EXCURSION FROM LINEARITY = $|i'_q(t) - o_{qa}(t)|_{max}$
mq
0
-mq
DITHER
+
$i'_q(t)$ +
Σ
Q
$\frac{1}{T_d}\int_0^{T_d} dt$
$o_{qa}(t)$
$m = \frac{\text{SAWTOOTH DITHER ZERO TO PEAK AMPLITUDE}}{\text{QUANTA SIZE q}}$

Fig. 6

line from P to the point $(0.5q, 0.5q)$. If $0.5 < R < 1.0$, locate point P at $[(R - 0.5)q, (R - 0.5)q + MEL_1]$ and proceed as above.

In Fig. 5, the curves for $m = 0.25$ and $m = 1.25$ have corresponding MEL's of $0.25q$ and $0.05q$, respectively, and their P points are located at $(0.25q, 0)$ and $(0.25q, 0.2q)$, respectively.

A comparably simple technique for the construction of the sinusoidal characteristic is not available.

For $0 < m < 0.5$, we have $OQG_\Delta = OQG_\sim = 0$ within the dead band $-(0.5 - m)q < i_q' < (0.5 - m)q$. However, with $0 < m < 0.5$ and $(0.5 - m)q < i_q' < (0.5 + m)q$, we obtain $OQG_\sim \neq OQG_\Delta = 0.5/m$.

For the sawtooth, if m lies in the range $0.5(n - 1) < m < 0.5n$, where integer $n \geq 1$, then $i_{qm}' = (0.5n - m)q$ for n odd and $i_{qm}' = [m - 0.5(n - 1)]q$ for n even. Therefore by finding $o_{qa}(i_q', m)$, setting $i_q' = i_{qm}'$ for the range of m in question, and setting $d/dm\,[o_{qa}(i_{qm}', m) - i_{qm}'] = 0$, the MEL_Δ maximizing value of m is found.

For example, for $0.5 < m < 1.0$ we get $i_{qm}' = (m - 0.5)q$. Consulting Table 2 and solving, we find that

$$(9) \qquad d/dm\,[o_{qa}(i_{qm}', m) - i_{qm}] = d/dm\,[1.5 - m - (0.5/m)]q$$

for $m = 1/\sqrt{2}$, which is the value of m at the maximum.

Table 2

SOME EXPRESSIONS FOR o_{qa}/q IN CLOSED FORM FOR SAWTOOTH DITHER

Definitions: $i_n' \triangleq i_q'/q$ and $s \triangleq i_n' + m$

Range of m	Range of s	Condition	o_{qa}/q
$0 \leq m \leq 0.5$	$0 < s < 0.5$		0
$0 \leq m \leq 0.5$	$0.5 < s < 1.5$	$i_n' - m \leq 0.5$	$(i_n' + m - 0.5)/(2m)$
$0 \leq m \leq 0.5$	$0.5 < s < 1.5$	$i_n' - m \geq 0.5$	1
$0.5 \leq m \leq 1$	$0.5 < s < 1.5$	$i_n' - m \leq -0.5$	i_n'/m
$0.5 \leq m \leq 1$	$0.5 < s < 1.5$	$i_n' - m \geq -0.5$	$(i_n' + m - 0.5)/(2m)$
$1 \leq m \leq 1.5$	$1 < s < 1.5$		i_n'/m
$1 \leq m \leq 1.5$	$1.5 < s < 2$		$(3i_n' + m - 1.5)/(2m)$

In a similar fashion, we find that the maxima of MEL_{Δ} for the ranges $1.0 < m < 1.5$ and $1.5 < m < 2.0$ occur at $m = \sqrt{3/2}$ and $m = \sqrt{3}$, respectively. For our purposes we can approximate by saying that the maxima of MEL_{Δ} occur at $m = 0$, 0.707, 1.225, 1.73, and at $m = 0.25 + 0.5(n + 3)$, where n is a positive integer.

D. Optimal Sawtooth Dither

A comparison of the plots for $MEL_{\sim}$ and MEL_{Δ} (Figs. 4 and 6) shows that not only is the sawtooth capable of perfect linearization, as has been mentioned, but that the sequence S_{Δ} of the maxima of MEL_{Δ} ($S_{\Delta} = 0.5q$, $0.085q$, $0.051q$, $0.035q$, $0.028q$, ...) converges to zero much more rapidly than the sequence $S_{\sim}$ of the maxima of $MEL_{\sim}$ ($S_{\sim} = 0.5q$, $0.17q$, $0.125q$, $0.105q$, $0.094q$, ...). Also, for all m, we have $MEL_{\Delta}(m) \leq MEL_{\sim}(m)$, the equality holding only for $0 \leq m \leq 0.45$. Therefore if the design procedure outlined for the sinusoidal dither is applied to the sawtooth for a given $m_{max} > 0.45$, then the sawtooth will be found to yield a smaller MEL than the sinusoid. Similarly, if the specifications call for a certain $MEL_{max} < 0.068q = MEL_{\sim}(m = 0.45)$, the design can be effected with a smaller value of m if a sawtooth rather than a sinusoid is employed.

Even though theory shows the sawtooth to be clearly superior to the sinusoid if values of $m_{max} > 0.45$ are permitted, practical considerations may obviate use of the sawtooth. As was mentioned earlier, injection of

dither directly into the quantizer may be physically impossible; dither may have to be injected together with the system input. Whereas $d_{\sim}(t)$ suffers only attenuation and phase shift as a result of operation A (if A in Fig. 2 is linear), $d_{\Delta}(t)$ so injected does not in general retain its sawtooth form. Only by performing the operation A-inverse (if this is possible) on $d_{\Delta}(t)$ prior to its injection in combination with the normal system input can one overcome this problem. Note that A-inverse is defined as that operation on any time function, say h(t), which when followed by operation A results in the function h(t).

It is evident that sinusoidal dither remains important for other than historical reasons.

4. SATURATING SAWTOOTH DITHER

In preceding portions of the present discussion, the normalized dither amplitude m was subject to the constraint that the quantizer must not be driven into saturation. As certain advantages accrue from employing dithers that are not subject to this constraint (i.e., saturating dithers), it is worthwhile to develop these properties numerically, as we shall now do for the sawtooth dither.

Consider a quantizer that saturates at its kth (k = 1, 2, ...) quanta level so that $|o_q| = kq$ for all $|i_q| > (k - 0.5)q$. Suppose that, of the class of sawtooth dithers, we wish to employ one that is theoretically capable of effecting perfect linearization, i.e., one for which m = 0.5n, where

n is a positive integer. Then the quantizer acting in conjunction with a low-pass or band-pass filter exhibits the following properties (see Appendix II for their derivation):

(1) For n = 1 (i.e., m = 0.5), we have $o_{qa} = i_q'$ everywhere in the linearized interval $i_q' < kq$. Outside of this interval, we have $|o_{qa}| = kq$. Let the quantizer's dynamic gain QDG $\triangleq \partial o_{qa}/\partial i_q'$, wherever this derivative exists. Hence for n = 1,

(10) (a) QDG = 1 for $|i_q'| < kq$,

(11) (b) QDG = 0 for $|i_q'| > kq$.

(2) Generalizing, for any integer n satisfying $1 \leq n \leq 2k$, where m = 0.5n, we have

(12) (a) QDG = 1 for $|i_q'| < [k + 0.5(1 - n)]q$,

(13) (b) QDG = (n - 1)/n for $[k + 0.5(1 - n)]q$

$< |i_q'| < [k + 0.5(3 - n)]q$,

(14) $QDG = (n - 2)/n$ for $[k + 0.5(3 - n)]q < |i_q'| < [k + 0.5(5 - n)]q$,

.
.
.

(15) $QDG = 1/n$ for $[k + 0.5(n - 3)q] < [k + 0.5(n - 1)]q$,

(16) $QDG = 0$ for $|i_q'| > [k + 0.5(n - 1)]q$.

(3) Let the maximum excursion from linearity when i_q' is allowed to range from zero to the saturation edge, where QDG becomes 0, be called $MEL_{\Delta s}$; i.e., $MEL_{\Delta s}(m = 0.5n) = |o_{qa}(i_q') - i_q'|_{max}$, where the maximizing value of $i_q' \triangleq i_{qm}' = [k + 0.5(n - 1)]q$. We find, for any integer n satisfying $1 \leq n \leq 2k$, that $MEL_{\Delta s}(m = 0.5n) = 0.5(n - 1)q$.

(4) We conclude from property (2) that the sensitive interval G, i.e., the interval for which $QDG \neq 0$, can grow only at the expense of the linear interval L (i.e., the interval for which QDG = 1), because $L + G = 4kq$ = twice the range which the quantizer output shows with no dither present.

As a consequence of properties (3) and (4), the engineer must compromise in the light of his particular performance specifications. He must decide whether to design for (1) maximum $L = 2kq$ = minimum G, and minimum

$MEL_{\Delta s} = 0$ by choosing $m = 0.5$; (ii) maximum $G = (4k - 1)q$, minimum $L = q$, and maximum $MEL_{\Delta s}(0.5n) = (k - 0.5)q$; or (iii) intermediate values of L, G, and $MEL_{\Delta s}$ (if $k > 1$).

The nature of this design trade-off can be seen in Fig. 7, where o_{qa} is plotted against positive i_q' for four values of $m = 0.5n$, and where the undithered quantizer output o_q saturates at $\pm 5q$ for all $i_q \gtrless \pm (k - 0.5)q$ (i.e., $k = 5$):

(a) For $n = 1$, $m = 0.5$, we have maximum $L = 10q =$ minimum G and $MEL_{\Delta s}(0.5) = 0$;

(b) For $n = 10$, we have maximum $G = 19q$, minimum $L = q$ and maximum $MEL_{\Delta s}(0.5n) = MEL_{\Delta s}(5) = 4.5q$;

(c) For $n = 2$, we have $L = 9q$, $G = 11q$, and $MEL_{\Delta s}(1) = 0.5q$; and

(d) For $n = 5$, we have $L = 5q$, $G = 14q$, and $MEL_{\Delta s}(2.5) = 2q$.

It is apparent from Fig. 7 that linearity here is a "currency" that can be "spent" to enlarge one's sensitive "frontage" G.

There is no theoretical limit on how large G (and $MEL_{\Delta s}$) can be made if one removes the condition $1 \leq n \leq 2k$ that has been imposed so as to make $L \geq q$.

5. CONCLUSIONS

If the raw quantizer output signal undergoes either a low-pass or a band-pass filtering (as it often does naturally) on its course to the system output, it is

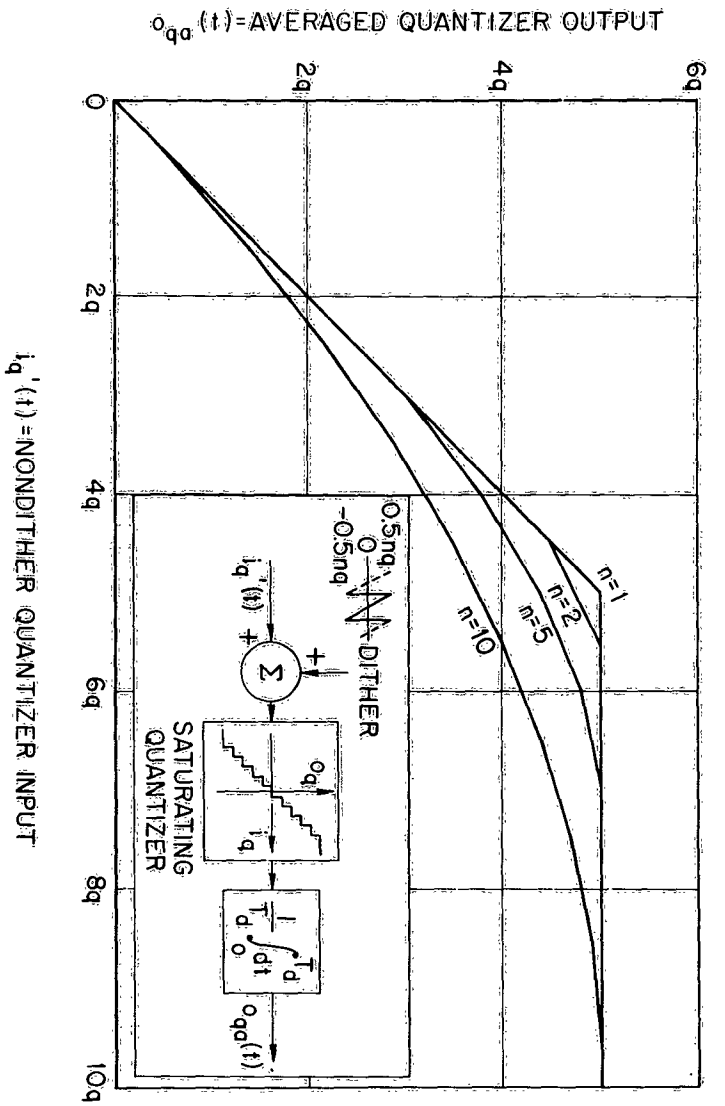
$o_{qa}(t)$=AVERAGED QUANTIZER OUTPUT
$i_q'(t)$=NONDITHER QUANTIZER INPUT
0
2q
4q
6q
8q
10q
n=1
n=2
n=5
n=10
0.5nq
0
-0.5nq
DITHER
$i_q'(t)$
Σ
SATURATING QUANTIZER
o_q
i_q
$\frac{1}{T_d}\int_0^{T_d} dt$
$o_{qa}(t)$

Fig. 7

possible in theory to alter the equivalent over-all quantizer gain OQG so that its deviation from a pure gain of unity is as small as is desired, if a high frequency dither can be injected. For a sinusoidal dither $d_{\sim}$, perfect linearization is approached in the limit as its normalized zero-to-peak amplitude m becomes large. A sawtooth dither d_{Δ} is capable, however, of effecting perfect linearization for values of m which are integral multiples of 1/2.

The OQG characteristic for either $d_{\sim}$ or d_{Δ} is completely specified if it is specified for nondither quantizer inputs which range one half a quantum in a amplitude, because of the periodicity and symmetry properties that the characteristic exhibits. The sawtooth OQG, OQG_{Δ}, being composed entirely of straight-line segments, is easily constructed, as is evident from the rules that we have presented for its construction.

The maximum excursion from linearity MEL(m) is specified for any m by comparing the OQG curve with the 45° line and finding the supremum of the absolute value of the difference between the two curves. It was found that the sinusoidal and sawtooth MEL(m) curves, i.e., the $MEL_{\sim}(m)$ and $MEL_{\Delta}(m)$ curves, and the extent of the corresponding dead-band regions are identical if $0 < m \leq 0.45$; but in the region $0.45 < m < 0.5$ the two curves part company because $MEL_{\sim}(m)$ rises from its first minimum at $m = 0.45$ while MEL_{Δ} continues to descend to its first minimum at

$m = 0.5$; $MEL_{\sim}(0.45) = MEL_{\Delta}(0.45) = 0.068q$ and $MEL_{\sim}(0.5) = 0.105q$ but $MEL_{\Delta}(0.5) = 0$.

Practical considerations place a bound on the maximum value which m is permitted to take; call this m_{max}. Then for $0 < m_{max} \leq 0.45$, $d_{\sim}$ and d_{Δ} are equally effective as linearizers. But for $m_{max} > 0.45$, d_{Δ} is preferable because $MEL_{\Delta} < MEL_{\sim}$ for $m > 0.45$. Similarly, if the design calls for an MEL < 0.068q then the sawtooth should be used. By employing the curves which have been presented, a method for finding the optimal dither amplitude (for either $d_{\sim}$ or d_{Δ}) in the face of design constraints has been proposed.

There is an advantage in employing a sawtooth if the dither amplitude is subject to slow amplitude drifting (e.g., because of ambient temperature variations) as one notes in the following example. Suppose that the design specifications call for a reduction of MEL to one tenth of its value without dither (i.e., from 0.5q to 0.05q), and that a sinusoidal dither has been decided upon because a sinusoidal signal generator is available. Then the smallest amplitude range for which this is possible is $m = 0.95 \pm 1\%$. On the other hand, a low amplitude sawtooth dither with $m = 0.5 \pm 10\%$ will also meet the specifications. The advantages of employing the smaller m with its greater tolerance should be apparent.

We have also considered the use of sawtooth dithers

with amplitudes so large as to drive the quantizer into saturation (it saturates at the $\pm$ kth quantization levels) as a means of extending the sensitive range of quantizer operation at the expense of incurring greater values of MEL. It was shown that it is possible to retain an interval of unity OQG around the origin of length q while the sensitive region of the quantizer is extended to $2[k + 0.5q\ (2k - 1)]q$. At the same time the MEL rises to $0.5(2k - 1)q$. The OQG characteristic for varying degrees of saturation was also developed.

Appendix I

PROOF THAT THE SAWTOOTH DITHER IS A PERFECT LINEARIZER IF m = n/2, WHERE n IS A POSITIVE INTEGER

Let $i_n' \triangleq i_q'/q$. Then it is necessary to prove that $o_{qa} = qi_n'$ only for $0 < i_n' < 0.5$, because of the symmetry and periodicity properties that have been mentioned in Section 3-A. For odd and even n, the maximum and minimum values that $o_q(t)$ can take on for a finite time interval with i_n' so bounded are the following:

Parity of n	Maximum $o_q(t)$	Minimum $o_q(t)$
even	nq/2	- nq/2
odd	q(n + 1)/2	- q(n - 1)/2

A. Even n. For n even, we have (2n + 1)/2 levels of $o_q(t)$. Examine $o_q(t)$ for the interval $0 < t < T_d$, noting that the duration for which o_q remains at any level remains constant as i_q' is varied slowly for all levels except the extreme ones, the maximum and minimum. For the nonextreme levels, the interval is T_d/n seconds. For the - nq/2 level, the interval is $W_1 = (T_d/n)(0.5 - i_n')$ seconds. For the nq/2 level, the interval is $W_2 = T_d/n - W_1 = (T_d/n)(0.5 + i_n')$ seconds. Therefore we have

$$(17) \qquad o_{qa} = (0.5nq/T_d)(W_2 - W_1)$$

$$= (0.5nq/T_d)(T_d/n)(0.5 + i_n' - 0.5 + i_n')$$

$$= qi_n' = i_q'$$

B. Odd n. For n odd, we also have $(2n + 1)/2$ levels of $o_q(t)$, and the duration is again T_d/n seconds for all levels except those at $q(n + 1)/2$ and $- q(n - 1)/2$. Let $W_1 \triangleq$ duration of the lowest level and $W_2 \triangleq$ duration of the highest level. Then we have $W_1 + W_2 = T_d/n$, but $W_1 = (T_d/n)(1 - i_n')$. Therefore we obtain $W_2 = T_d i_n'/n$,

$$o_{qa} = (q/T_d) \left\{(n - 1)/2 \left[T_d/n - W_1\right] + \left[(n + 1)/2\right] W_2\right\} \tag{18}$$

Substitution yields

$$o_{qa} = q i_n' = i_q' \tag{19}$$

Appendix II

DERIVATION OF THE PROPERTIES ASCRIBED TO SATURATING SAWTOOTH DITHER

The object of this appendix is the derivation of properties (1), (2), and (3) which have been ascribed to the saturating sawtooth dither in Sec. 4.

With regard to property (1), observe that it is a special case of property (2) with n = 1.

Property (2) is composed of two parts, (a) and (b), which will now be derived:

(a) Quantizer saturation does not occur unless the peak of i_q satisfies $i_q = i_q' + 0.5nq > (k + 0.5)q$. Then (a) applies to nonsaturating values of i_q (as proven in Appendix I), and the linearization is perfect, so that QDG = 1.

(b) Examine $o_q(t)$ for the interval $0 < t < T_d$. Call the interval of length q for which Eq. (13) applies the saturating interval $1 = SI_1$, the interval for which Eq. (14) applies SI_2, and so on. There will be $n - 1$ such intervals for which ODG $\neq 0$, because only at the lower edge of SI_n do all values of $i_q(t)$ lie in the saturation region; i.e., we have $i_q(t) > (k + 0.5)q$.

For SI_1, the minimum level of $o_q(t)$ is $(k - n + 1)q$, and it has a duration $W_1 = (T_d/n)\,[1 - (i_q' + B)/q]$ seconds, where $B = [k + 0.5(n - 1)]q$. The maximum level of $o_q(t)$ is, of course, the saturation level kq, and it

has a duration of $W_2 = [1 + (i_q' - B)/q]$ seconds. The durations of the levels lying between these two extremes do not change for any saturating interval. Consequently these levels can be ignored in solving for

$$(20)\quad QDG = (n-1)/n\ \partial o_{qa}/\partial i_q'$$

$$= q/T_d\ [k(\partial W_2/\partial i_q') + (k-n-1)(\partial W_1/\partial i_q')].$$

For QI_2, the minimum level of $o_q(t)$ is $(k-n+2)q$, and its duration is

$$(21)\quad W_1 = T_d/n\ [1 - (i_q' - B - q)/q] \text{ seconds;}$$

the maximum level is again kq, and its duration is

$$(22)\quad W_2 = T_d/n\ [2 - (i_q' - B - q)/q] \text{ seconds.}$$

Therefore we have

$$(23)\quad QDG = q/T_d\ [k(\partial W_2/\partial i_q') + (k-n+2)(\partial W_1/\partial i_q')] = \frac{n-2}{2}$$

For the jth quantizing interval $1 \leq j \leq n$, the minimum level is $(k-n+j)q$, and its duration is

(24) $W_1 = T_d/n\ [1 - \langle i_q' - B - (j-1)q \rangle/q]$ sec. and

$$\partial W_1/\partial i_q' = -1/q.$$

The maximum level is kq, and its duration is

(25) $W_2 = T_d/n\ [j + \langle i_q' - B - (j-1)q \rangle/q]$ sec. and

$$\partial W_2/\partial i_q' = 1/q.$$

Therefore we obtain

(26) $QDG = T_d/n(q/T_d)[k/q - (k - n + j)/q] = \frac{n-j}{n}$. Q.E.D.

Property (3) follows directly from the fact that QDG decreases monotonically, with the result that the divergence between the line $o_{qa} = i_q'$ and the $o_{qa}(i_q')$-vs-i_q' characteristic increases monotonically. The maximum magnitude of this divergence, i.e., MEL, is obtained by setting i_q' equal to its greatest permissible value under the definition, i.e., $i_q' = [k + 0.5(n-1)]q$.

14. Oldenburger, R., and R. C. Boyer, Effects of Extra Sinusoidal Inputs to Nonlinear Systems, Winter Annual Meeting ASME, New York, November 1961.

15. Gibson, J. E., and R. Sridhar, A New Dual-Input Describing Function and an Application to the Stability of Forced Nonlinear Systems, Joint Automatic Control Conference New York, June 1962, paper 9-1.

16. Vander Velde, W. E., and A. Gleb, The Dynamics of Limit Cycle Amplitude Regulation, Joint Automatic Control Conference, New York, June 1962, paper 7-4.

17. Ishikawa, T., Linearization of Contactor Control Systems by External Dither Signals, Technical Report No. 2103-2, Stanford Electronics Laboratories, October 1960.

AD-296 598

PLEASE INSERT THE ATTACHED ERRATA SHEET IN YOUR COPY OF RM-3271-PR, WHICH WAS SENT TO YOU RECENTLY.

THE RAND CORPORATION

Errata Sheet for RM-3271-PR

REMOVING THE NOISE FROM THE QUANTIZATION PROCESS BY DITHERING: LINEARIZATION

G. G. Furman

1. Page 10, equation (2) should read:

$$o_{qa} = (1/T_d) \int_0^{T_d} o_q(t)\, dt.$$

2. Page 15, 12th line from the bottom of the page should read:

occur at $m \cong 0.15 + 0.5n$, where integer $n \geq 1$, i.e., at

3. Page 35, second line from the bottom of page, the right square bracket is not complete. The right bracket should be completed and the expression should read:

..... $B = [k + 0.5(n - 1)]q$.

The following is a list of captions for the figures.

Fig. 1 — The Periodic Quantization Type of Nonlinearity

Fig. 2 — Problem Formulation

Fig. 3 — Over-all Quantizer Gain for Sinusoidal Dither

Fig. 4 — Maximum Excursion from Linearity for Sinusoidal Dither

Fig. 5 — Over-all Quantizer Gain for Sawtooth Dither

Fig. 6 — Maximum Excursion from Linearity for Sawtooth Dither

Fig. 7 — Over-all Quantizer Gain for a Saturating Sawtooth Dither